话题检测与跟踪研究进展综述

张瑾 杨森 王孝宗 罗维 杜攀 程学旗

摘要：随着互联网信息的指数增长，为了提高信息挖掘的效率，信息检索与话题检测等技术近年得到了广泛关注。本文首先回顾了话题检测与跟踪技术发展的历史，并在介绍传统话题检测方法的基础上，从突发性检测与基于社会网络的话题检测与跟踪方法两个方面进行深入探讨；对话题检测与跟踪的评价方法进行了分析；最后展望了话题检测与跟踪方法的发展趋势。

关键词： 话题检测与跟踪，突发检测，社会网络；

1 引言

随着互联网技术的蓬勃发展和广泛普及，网络上的信息量呈指数增长，信息过剩与知识匮乏并存的矛盾日益凸显。浩瀚的网络数据远远超出了人们的掌控能力，因此，如何有效地组织并展现 Web 数据，提高知识获取的效率，长期以来一直是一个热点研究问题。借助于话题检测与跟踪技术可以把信息按主题分类组织，将特定时间段内最活跃的话题智能地推送给用户，并按照用户的需求跟踪话题的动态演化过程，从而为用户有效掌握社会动向和重大事件提供极大便利。尤其是面向热点话题与突发话题的相关应用更得到了广泛的关注。

同时，随着 Web 2.0 的应用与发展，社会网络变得越来越普及。与以往的新闻网络媒介不同，社会网络更加强调用户的参与性。如果能够有效地在社会网络上自动检测和跟踪话题，无疑能够方便用户在社会网络上寻找并全面了解其所感兴趣的事件或者话题。然而由于社会网络上的数据主要由普通用户产生，这些数据无论是用词、形式还是具体内容的质量都参差不齐，给话题检索带来很大困难。值得注意的另一方面是，用户的广泛参与，为话题检测和跟踪提供了可利用的新的数据信息。社会网络上的话题检测的数据不仅局限文本信息，还可以利用非文本信息。这些新特点使面向社会网络的话题检测和跟踪方法的研究在最近几年得到了重点关注。

在本文中，我们首先回顾了话题检测与跟踪的历史；在介绍传统话题检测方法的基础上，结合我们的研究成果从突发性检测与基于社会网络的话题检测与跟踪方法两个方面进行深入探讨；对目前的话题检测与跟踪的评价方法进行了分析；最后对话题检测与跟踪方法未来的发展趋势进行了展望。

2 研究现状

话题检测和跟踪研究已经开展十多年了。在现有的研究中，话题被定义为某个事件或活动及所有与其相关的事件或活动，而事件则定义为在某个特定的时间或地点发生的某件独特的事情[4]。在以往的研究中，事件和话题的定义差别微小并且经常可以互换。话题的检测可以分为两个相对独立的子任务，即历史话题检测（或回顾式话题检测）和在线话题检测。历史话题检测是指在已知所有的检测数据后，在该数据集上检测其中隐含的所有话题。在线话题检测是指在进行话题检测的时候，检测数据只是部分可知，并且新的数据是以在线的形式不断地呈现给检测系统，要求话题检测系统能够即时地对当前新到达文本进行话题的判断，即判断当前新文档是新的话题还是属于某个已有的历史话题。话题的跟踪任务是指对于一个

事先指定的话题（按照某种形式呈现），在在线数据输出模式中，在新数据到达之前判别当前文档是否属于该指定的话题。

在话题检测与跟踪（Topic Detect and Tracking, TDT）评测中，用于进行话题检测和跟踪的语料为新闻数据，这些数据包括新闻文本和转录语料，通常按照时间有序排列，并且目标事件已经被人工标注。在评测中，话题检测和跟踪研究[1,4,22,24,25]被进一步划分为三个子任务：数据流的切分、事件的检测、事件的跟踪。数据流的切分任务被定义为对连续的文本数据流按照报道内容进行切分，正确识别出与相邻报道的边界。事件的检测可以进一步被划分为历史事件检测（Retrospective Event Detection, RED）和在线新事件检测（Online New Event Detection, NED）[4]。历史事件检测是指在一个给定的报道集合中找出所有隐含的事件，其任务就是对目标数据集进行聚类，每一个聚类结果簇表示一个事件。而在线新事件检测的目标是以在线的方式在报道流中识别新事件。当有新报道到达时，要求在线新事件检测方法能够对该报道进行分析并且在下一个报道到达之前判断该报道是否讨论了一个新的事件。而事件的跟踪是指在新到达的报道中找出所有与已知事件相关的报道。

由于我们的研究主要集中在话题的检测和跟踪上，所以下面我们将主要分析已有的事件检测和跟踪方法的研究而忽略数据流的切分研究。关于话题检测和跟踪的研究可以从方法上划分为两类。第一类方法主要是寻找适合于话题检测和跟踪的新的聚类算法或者对已有的聚类算法进行改造。另一类方法则集中于挖掘新的话题特征来提高检测和跟踪的效果。值得注意的是，在有些研究中，比如引文[1]等，这种划分有时候并不明显。为了简便起见，我们不再一一进行严格的说明。

3　主要方法

话题检测与跟踪系统的主要工作是准确地检测话题并跟踪话题的动态演化过程，其中最关键的问题是如何进行话题的检测。艾伦（James Allan）等人[4]将话题检测分为两个分支：一个是回顾式话题检测，即对语料库中的文档以话题为单位进行再组织，本质上是一个无指导的分类问题，把讨论同一个话题的文档划分到一起；另一个是在线新事件话题检测，指面对增量式到来的在线文档流，顺次处理，同时决定它是属于某个已经标注的话题还是讨论了一个新的话题。在线话题挖掘与回顾式话题挖掘的主要区别在于：在线话题挖掘面对的是增量式的文档流，而回顾式话题挖掘面对的是整个文档语料库。

3.1 话题表示及相似度度量

所谓“表示”就是指将文档和话题抽象成计算机可计算、可比较的模型。相似度度量包括计算文档与文档之间、文档与话题之间以及话题与话题之间的相似性。这两个问题是高度相关的，每个表示模型对应了一种或多种相似度计算方式。

常用的话题表示模型主要是向量空间模型、概率检索模型和语言模型。

向量空间模型（Vector Space Model，VSM）：该模型将文档和话题都表示成一个向量，向量的每一维表示一个词。这样整个词典构成了空间中的所有维，每个文档和话题变成了空间中的一个向量（点）。

与向量空间模型对应的一种最自然的相似度度量方法就是计算向量的余弦。即

$$\boldsymbol{sim}(\vec{\boldsymbol{d}}_i,\vec{\boldsymbol{d}}_j)=\boldsymbol{cos}(\vec{\boldsymbol{d}}_i,\vec{\boldsymbol{d}}_j) \qquad (1)$$

概率检索模型（Probability Retrieval Model）：该模型也将文档和话题表示成一个词集，将相似度看作一个概率值，即给出一个查询（Query），文档（Document）与该查询相关的概率。与概率模型对应的一种常用的相似度度量方法是 BM25 公式：

$$sim(D,Q)=\sum_{i=1}^{n} IDF(q_i)\times\frac{f(q_i,D)\times(k_1+1)}{f(q_i,D)+k_1(1-b+b\times\frac{|D|}{avgdl})} \quad (2)$$

$$IDF(q_i)=log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

其中，k_1，b 为自由参数，q_i 是组成查询 Q 的单词，$f(q_i,D)$ 表示词 q_i 在文档 D 中的词频，$|D|$ 表示文档 D 的长度，$avgdl$ 表示语料库中文档的平均长度。

语言模型（Language Model）：统计语言模型认为语言就是字母表上的一种概率分布，特征集合 $\{t_i|i=1,\cdots,n\}$ 在某个文档 D 中形成一个分布，这个概率分布称为一个语言模型。语言模型将每个文档作为一个语言模型，整个语料集也是一个语言模型。在语言模型中，计算查询与文档的相关性定义为从一个语言模型生成另一个语言模型的概率 $P(Q,D)$。常用的是库尔贝克-莱布勒(Kullback-Leibler) 距离，其计算公式为：

$$KL(Q,D)=\sum_{w_i\in Q} P_Q(w_i)\,log\frac{P_Q(w_i)}{P_D(w_i)} \quad (3)$$

由于语言模型的稀疏性，有可能出现零概率词，因此必须解决对零概率词平滑性的问题。

3.2 话题检测方法

关于话题检测和跟踪的研究方法可以分为两类：第一类旨在寻找适合于话题检测和跟踪的新的聚类算法或者对已有的聚类算法进行改造；第二类则集中于挖掘新的话题特征来提高检测和跟踪的效果。本节，我们将从这两种类别来简述已有的话题检测和跟踪方法。

3.2.1 改进聚类算法

从话题检测的定义来看，话题检测和跟踪跟聚类算法的研究具有较大的相似性。因此，人们便试图寻找更加适合于话题检测和跟踪的聚类算法。

杨一民等人在1998年提出了一个基于平均分组的层次聚类（Group Average Clustering，GAC）的历史事件检测方法[1]。GAC是一个凝聚式的聚类算法，它的目标是使结果簇中文本对的平均相似度最大化。杨一民等人在GAC的基础上，提出了切分和重聚类的方法。该方法通过对数据的切分充分利用了事件聚集性特征，即同一事件的报道倾向于聚集在一个相对较小的时间区域，并且在重聚类的时候能够将初始分块边界对聚类结果的影响降到最低。

艾伦等人在基于K-平均值（K-Means）聚类算法的基础上提出了多重K-平均值方法来进行历史事件的检测[4]。多重K-平均值的基本思想为：在每个给定的时间点，已知有k个簇，对于每个报道，找到与其最近的簇。如果该距离小于某个阈值，那么将该报道划分到该簇中；如果该距离大于某个阈值，则由该文档生成一个新的簇，在此基础之上再进行常规的K-平均值聚类算法。

另外，李志伟等人（音译， Li et al. ）在2005年提出了基于概率模型的历史新闻事件检测方法[2]。该方法使用概率生成模型结合内容和时间信息进行历史事件检测。对于每个已知的事件用一个概率生成模型来表示。对于每一个文档，找到生成该文档的概率最大的那个生

成模型所代表的事件，则该文档即属于该事件。同时，该方法还考虑到对同一个事件的报道往往会分布在多个新闻源上。所以，在该方法中，结合了不同数据源在相同时间报道的具有较为相似的事件来帮助进行事件的检测。

引文[23]提出了一种基于多策略优化的分治聚类算法。该算法能够首先将全部的数据分为具有一定的相似性的分组，然后对各个分组分别进行聚类，得到每个分组内部的聚类结果，即"微类"。在此基础上，再对所有的微类进行聚类，得到最终的结果话题。同时，在聚类的过程中，该方法采用了多种策略的优化方式来改善聚类的效果。

需要说明的是，虽然上面介绍的方法在初始引入的时候只是针对历史事件检测或者只是针对在线话题检测，但是，由于历史事件检测本身在某种程度上能够分解为在线话题的检测，所以，上面介绍的在线话题检测的方法大都可以运用于历史事件的检测任务。下一节的介绍中，我们将不再区分所介绍的方法所具体针对的检测任务。

3.2.2 挖掘话题特征

另外一种研究思路是挖掘话题所固有的特征来改进话题检测和跟踪的效果。话题特征包括话题的时间聚集性、话题的特征词、话题的生命演变特征以及话题的命名实体等。

目前较广泛采用的一种思路是利用话题的各种特征来寻找合适的途径控制话题检测和跟踪时的话题阈值，旨在寻找能够兼容相对较为广泛的阈值设定方法。该方面的研究包括了艾伦等人在引文[4]中提出的时间惩罚策略、周子铨等人在引文[6]中提出的增量式概率浅层语义索引(Probabilistic Latent Semantic Indexing，PLSI)在线事件检测算法、陈致杰等人在引文[7]中提出的基于隐马尔科夫模型的事件生命特征识别方法。

艾伦等人主要是使用单遍法（Single Pass）聚类算法和一个新的阈值控制模型来进行在线新事件的检测。该阈值控制策略的基本思想是：相距较远的两个报道必须具有较大的相似性才能将其划分为同一个事件，而相距较近的两个文档则需要较小的相似性将其归为同一个事件。周子铨等人实现的增量式概率浅层语义索引模型主要目的是扩大检测阈值设定的有效范围。相对于基于向量空间模型的文本和话题表示模型，概率浅层语义索引由于能够更加有效地表示话题，因此可以容纳更为宽广的阈值范围。陈致杰等人的生命特征识别方法认为事件的发展具有一定的特定模式，即产生、发展、壮大和消亡，因此可以通过隐马尔科夫模型训练已知的若干种具有不同生命特征的事件演变方式，然后再对每个新事件的行为模式进行预测。通过对不同的事件演变阶段赋以不同的检测阈值，即通过动态话题阈值策略来改进已有话题检测方法的检测效果。引文[3]在2004年提出了一种将文本中的特征词进行分类的方法，即将关键词分为地点、名字、时间和一般特征词等，然后在各自的类别上进行文本内容的比较。引文[8]在2004年提出了通过文本分类和命名实体来改进新事件检测的效果。该论文通过对文本进行分类，对不同类别给予不同的相似性阈值，通过文本的多重表示方法，即将一个文本表示为三个部分：由所有的特征词构成的表示、由命名实体构成的表示、由非命名实体构成的表示，改进文本内容相似性计算的效果。

在中文研究领域，引文[21]首先对文本特征进行分类，将所有的词特征分为人名、地名和主题信息等，并对于每种类别赋予不同的相似性比较系数。在此基础上，将每个特征词的权重定义为该特征词的词频与其所属类别的相似性比较系数的乘积。该方法通过给予不同类别的特征词以不同的权重计算系数，可以加强特定类别在文本相似度计算中的权重，从而提升话题检测的精度。此外，引文[22]提出了一种通过构建地理树的方法来计算命名实体的相似性。由于在地名的表述中，不同的地名可能共享某种程度的相似性，因此引入预先设定的地理树能够较为有效地解决不同地名之间共享相似度的问题。但是，该方法仅局限于地名的比较，对于其他词性，其应用局限性较大。

3.3 突发事件检测

近些年，突发事件的研究越来越引起人们的注意。突发特征是指伴随着事件的发生，若干与该事件密切相关的某种特征，比如文档或者词语等会出现反常爆发的特性。突发事件就是指具有突发特征的事件。目前，关于突发事件的研究主要集中于从数据集中寻找所有与该事件关联的突发词，然后再将这些突发词进行组合以形成该突发事件的特征，从而用突发特征描述突发事件。这类研究的目标与传统的话题检测与跟踪中的事件检测有所不同。突发事件的研究旨在通过一组突发词来识别出一个突发事件。而在话题检测与跟踪中，事件是通过文档集来表示的。突发事件的研究不再局限于新闻数据，也包括了查询日志、邮件、博客等语料。同样，突发事件的检测也分为两类，即历史突发事件的检测和在线突发事件的检测。

在2002年，引文[10]提出了一个简单的、功能强大的用于文本流突发检测的自动机模型。该自动机模型通过自动机来模拟特征词的状态及状态之间的转换。不同的状态表示了词的不同的出现频率，而这些状态间的转换则表示了突发的产生或者消亡。通过给予状态转换以一定的惩罚，自动机模型能够有效地防止错误检测出过多的非突发性词频变化。在该文章中，作者将自动机模型应用到邮件集和新闻集中，并且证明了自动机方法进行突发检测的有效性。

引文[11]提出了一个无参数的突发事件检测方法。和自动机模型不同的是，该方法在可以有效地在文本流中自动检测突发事件的同时不需要用户指定任何参数。同样，该方法的目标也是寻找突发词集，而每个突发词集就表示了一个突发事件。具体来讲，二项式分布被用来表示一个词在文本中出现的可能性。通过该分布，一个可能性非常小的词的出现频率突然增长就被认为是该词的突发。并且通过在新闻数据上的实验分析，证明了该方法的有效性。

引文[12]提出了一种通过离散傅立叶变换将时序信号分解为一系列正余弦信号，分析其中能量值（傅立叶系数）最大的信号的行为，来识别非周期性突发词和周期性突发词的方法。该方法还能够分别识别弱突发性词和强突发性词。该方法能够较好地处理周期性突发事件的识别，同时还能够较好地识别弱突发性事件。

引文[13]通过对自动机模型的改造，得到一个在线的突发事件检测方法。该方法能够较好地在查询日志中以在线的方式检测突发事件。由于采用动态规划来求解当前的最优状态，在内存中每个时刻只需要很小代价来保持上一个时间点的每种状态值。

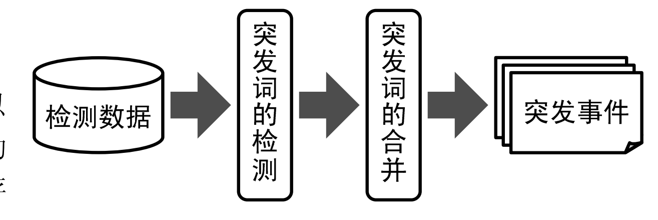


图1. 突发事件检测的一般流程

引文[14]将突发事件的检测思想进一步应用到识别博客中的突发用户群。

综上所述，突发事件检测的研究包含两个基本的步骤，即突发词的识别和突发词的合并。突发词的识别旨在检测出数据集中所有的具有突发特征的特征词。突发词的合并立足于用这些突发特征词构建最终的突发事件特征。突发事件检测的流程如图1所示。

3.4 基于社会网络的话题检测和跟踪

与以往的新闻网络媒介不同，社会网络更加强调用户的参与性。并且，由于社会网络给用户提供了一个方便的信息交流平台，各种各样的具体形式的网络媒介得到了很大的发展。比如：博客、网络论坛、社交网络、视频共享网站以及最近兴起的微博等。如果能够有效地在社会网络上自动检测和跟踪话题，无疑能够方便用户在社会网络上寻找并全面了解其所感

兴趣的事件或者话题。

以往关于话题的研究，特别是话题检测与跟踪的研究集中在新闻数据上。而新闻数据形式严谨，用词确切，内容具体。这些都与社会网络上的数据特征有很大差别。由于社会网络上的数据主要由普通用户产生，这些数据无论是用词、形式还是具体的内容，质量都没有保证；另外，用户的广泛参与也为话题检测和跟踪提供了可利用的新的数据信息，也就是说，社会网络上的话题检测不仅局限于文本信息，还可以利用非文本信息。这两个特点使我们有必要寻找新的更加适合于社会网络的话题检测和跟踪的方法。

虽然关于社会网络上的话题检测和跟踪的研究具有很大的价值。但是，由于数据质量参差不齐，要得到有效的话题检测和跟踪算法并不是一件十分容易的事情。并且，不同的社会网络形式也会对话题检测和跟踪方法产生较大的影响。有关的话题检测和跟踪研究几乎涉及到各种形式的社会网络数据，包括查询日志、博客、网络论坛、视频共享平台等。随着各种新型的应用社会网络平台的出现，话题检测和跟踪的方法也需要不断地进行改进。

朱明亮（音译，MingLiang Zhu）等人在2008年提供了一个在主题讨论社区（Threaded Discussion[1] Community）中检测和跟踪话题的方法[16]，该研究集中于设计有效方法来消除潜在的噪声的影响，并且通过引入用户的相似性来改进线索（thread）相似性计算的效果。刘路（音译， Liu，Lu）等人对视频话题检测的方法进行了研究[17]，通过视频和标注词形成二部图，然后再在该二部图上通过联合聚类（Co-clustering）算法[18]进行话题检测与跟踪。在YouTube上的实验分析表明，该方法能够较好地在视频网页上检测和跟踪话题。在2007年，N. 班塞尔（Nilesh Bansal）等人通过对用户查询的分析在文本数据流上检测事件。该方法的基本思想是：首先，从查询日志中找出具有突发性的查询词。然后，使用这些突发词的查询结果构建事件。引文[21]通过对用户查询的分析在用户产生的数据（User Generated Content, UGC）流上检测事件。该方法的基本思想是：首先，从查询日志中找出具有突发性的查询词。然后，使用这些突发词的查询结果构建事件。该文章使用查询日志和博客数据进行了实验并且取得了较好的事件检测效果。

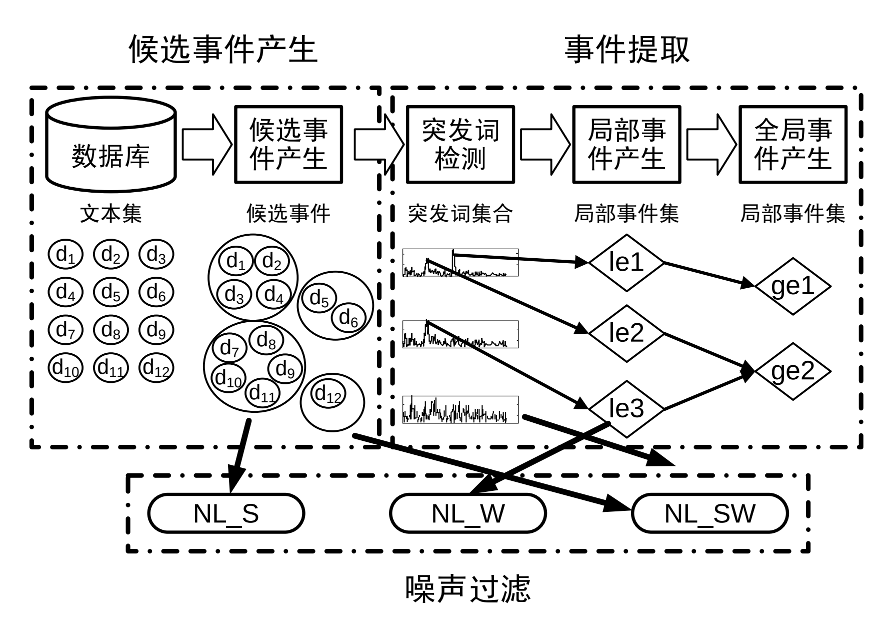


图2. 带噪音过滤的突发事件检测框架

近来，我们对论坛上的话题检测进行了较为深入的研究。引文[19]提出了论坛上的突发

[1] 此概念尚无统一译法，亦有译作“按线索讨论”、“基于线索的讨论”、“穿插讨论”等。

话题的检测方法，首先利用突发性特征来过滤具有突发性的特征词和用户，再通过对突发词的组合来构建突发话题。同时，通过突发用户群来进一步验证所检测到的突发话题。另外，我们注意到论坛上的帖子质量很难得到保障：一方面，由于用户产生的文本内容本身质量参差不齐；另一方面，论坛上存在大量的非事件文本。在实际的话题检测过程中，这些噪音都会对结果产生较大的影响。为了过滤上述论坛噪音，我们结合了文本内容相似性和事件的突发性来进行论坛上的话题检测[20]，并在此基础上提出了一种基于噪音数据的突发事件检测框架，如图2所示。

我们对收集的腾讯论坛数据进行人工标注，并使用该数据集对我们的方法进行了评测。通过实验发现，我们所提出的带噪音过滤的突发事件检测方法能够较好提升已有方法在噪音数据上的检测效果。

3.5 话题检测和跟踪的评价指标

对于话题检测和跟踪的评价，一般采用多种评估标准。这些标准包括：准确率（Precision, p）、召回率（Recall, r）、F值（F_1）、误报率（False Alarm Rate, $False$）、漏报率（Miss Rate, $Miss$）、归一化的检测代价（Normal Cost Value, $(C_{det})_{norm}$）、以及相应的宏平均值和微平均值等[8,9]。

根据已有的话题检测和跟踪的评测方法，话题检测算法的评价方法为：对于检测算法检测出的任意数目的结果话题，对该检测和跟踪算法的评测只集中在所预先人为选定并且标注的若干话题上。对于标注话题的评测，我们在检测的结果话题中，找出与标准话题具有最大公共文档集的结果话题，作为该标注话题的对应检测结果[1]。所谓的公共文档集是指评测话题和某个结果话题的共享文档集。该评价方法适合于历史话题检测的性能评测。同时我们也可以将该评测使用在在线的话题检测的性能评测上。

各种评测的具体指标则基于关联矩阵（Contingency Matrix）来获得，关联矩阵中各项的值代表了满足该项要求的文档的数目，如下表所示。

表 1 话题检测结果的关联矩阵

	在标注话题中	不在标准话题中
在检测结果话题中	a	b
不在检测结果话题中	c	d

基于该关联矩阵，准确率、召回率、F值、误报率、漏报率的定义分别为：

- $p = a/(a+b)$ （$(a+b) > 0$，否则 p 为未定义）；
- $r = a/(a+c)$ （$(a+c) > 0$，否则 r 为未定义）；
- $F_1 = 2pr/(p+r) = 2a/(2a+b+c)$ （$2a+b+c > 0$，否则 F_1 未定义）；
- $False = b/(b+d)$ （$(b+d) > 0$，否则 $False$ 为未定义）；
- $Miss = c/(a+c)$ （$(a+c) > 0$，否则 $Miss$ 为未定义）；

准确率是指在在检测出的结果事件中，真正属于该事件的文档所占的比例。召回率是指检测出的事件的文档与标准事件的文档的比例。由于准确率和召回率之间经常是互为消长的关系，为了获得比较高的准确率通常要牺牲召回率，同样为了获得比较高的召回率通常要牺牲准确率。只用一种评价指标可能会导致错误的评价结论，一种较好的方法是把准确率和召回率进行统一考虑，常用的方法是使用 F_1 值。

误报率是指在检测的结果事件中的不属于标注话题的文档与所有不属于标注话题的文档比例。漏报率是指标注话题中未能检测出的文档比例。与正确率和召回率的关系类似，误报率和漏报率也是互为消长的关系。所以，需要一种更好的能够将这两种评价指标融合的指标，即检测代价。

检测代价结合了误报率和漏报率，其定义为：

$$C_{det} = Miss \times P_{miss} \times P_{target} + false \times P_{false} \times (1 - P_{target}) \quad (4)$$

其中，P_{miss} 和 P_{false} 分别表示漏报和误报的条件概率，P_{target} 是一个先验的概率。C_{det} 越小就表示算法的检测效果越好。然而，由于 C_{det} 的定义与先验概率有关，为了更好地表示检测算法的性能，在话题检测中，更常用 C_{det} 的归一化值，即 $(C_{det})_{norm}$，其定义为[9]:

$$(C_{det})_{norm} = \frac{C_{det}}{min((P_{miss} \times P_{target}),(P_{false} \times (1 - P_{target})))} \quad (5)$$

可见，$(C_{det})_{norm}$ 值不超过1。同样，$(C_{det})_{norm}$ 越小表示检测算法的检测效果越好。在我们的实验中，P_{target} 被设定为0.02。

由于话题检测的评价所使用的评测事件的数目一般大于 1。所以，为了描述检测算法对每个评测事件的综合检测效果，我们还需要使用宏平均和微平均。宏平均是指直接在各个评测事件的评价指标上进行加权平均，而微平均是指首先将各个评测事件的关联矩阵相加，然后在总的关联矩阵上计算总的评测指标[5]。

4 总结

虽然话题检测与跟踪研究已经开展很多年，但是由于互联网数据来源的多样性与特征抽取的不确定性带来的困难，目前话题检测的研究主要集中于新闻类数据上，社会网络上话题检测的研究相对较少。随着社会网络的兴起，特别是论坛、微博等的广泛应用，面向突发事件等特定需求和面向社会网络数据的新兴应用的话题检测具有越来越重要的意义。我们相信，随着面向社会网络数据的特征选择方法以及用户行为与文本内容关联挖掘方法的综合应用，话题检测与跟踪技术的研究与应用会得到进一步的发展。

参考文献

[1] Y. Yang, T. Pierce, and J. Carbonell, A study on retrospective and online event detection. In proceedings of the 21th annual international ACM SIGIR conference on research and development in information retrieval, pages 28-36, 1998.

[2] Zhiwei Li, Bin Wang, Mingjing Li, and Wei-Ying Ma, A probabilistic model for retrospective news event detection. In proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, pages 106-113, 2005.

[3] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. Simple Semantics in Topic Detection and Tracking. Kluwer Academic Publishers, pages 347-368, 2004.

[4] Jame Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study: Final Report. In proceedings of DARPA broadcast news transcription and understanding workshop, pages 194-218, 1998.

[5] Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto. Modern information retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999.

[6] Tzu-Chuan Chou, Meng Chang Chen. Using Incremental PLSI for Threshold-Resilient Online Event Analysis, IEEE Trans. Know. Data Eng. 20, 3, pages 289-299, 2008.

[7] Chien Chin Chen, Meng Chang Chen, Mimg-syan Chen. An Adaptive Threshold Framework for Event Detection Using HMM-Based Life Profiles. ACM Transactions on Information Systems, 2009.

[8] Giridhar Kumaran, and James Allan. Text classification and named entities for new event detection. In proceedings of the seventeenth annual international ACM SIGIR conference on research and development in information retrieval, pages 297-304, 2004.

[9] J. Allan, V. Lavrenko, and R Swan. Explorations within topic tracking and detection. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic: Massachusetts, pages 197-224, 2002.

[10] J. Kleinberg, Bursty and hierarchical structure in streams. In proceedings of 8th ACM SIGKDD international conference on knowledge discovery and data mining, pages 373-397, 2002.

[11] Gabriel Pui Cheong Fung, Jeffery Xu Yu, Philips S. Yu, Hongjun Lu. Parameter free bursty events detection in text streams. In proceedings of the 31st international conference on very large data bases, pages 181-192, 2005.

[12] Qi He, Kuiyu Chang, and Ee-Peng Lim. Analyzing feature trajectories for event detection. In proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pages 207-214, 2007.

[13] Nish Parikh, Neel Sundaresan. Scalable and near real-time burst detection from eCommmerce queries. In proceedings of 14th ACM SIGKDD international conference on knowledge discovery and data mining, pages 972-980, 2008.

[14] Nilesh Bansal, Nick Koudas. BlogScope: a system for online analysis of high volume text streams. In proceedings of the 33rd international conference on very large data bases, September 23-27, 2007, Vienna, Austria.

[15] Meishan Hu, Aixin Sun, and En-Peng Lim. Event detection with common user interests. In proceeding of the 10th ACM workshop on Web information and data management, pages 1-8, 2008.

[16] MingLiang Zhu, Weiming Hu, Ou Wu. Topic detection and tracking for threaded discussion communities. In proceedings of the 2008 IEEE/WIC/ACM international conference on Web intelligence and intelligent agent technology, pages 77-83. 2008.

[17] Lu Liu, Lifeng Sun, Yong Rui. Web video topic discovery and tracking via bipartite graph reinforcement model. In proceeding of the 17th international conference on World Wide Web, pages 1009-1018, 2008.

[18] Inderjit S. Dhilon, Subramanyam Mallela, Dharmendra S. Modha. Information- theoretic co-clustering. In proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pages 89-98, 2003.

[19] You Chen, Sen Yang, Xueqi Cheng. Bursty topics extraction for web forums. In proceeding of the eleventh international workshop on Web information and data management, pages 55-58, 2009.

[20] Sen Yang, Xueqi Cheng, You Chen, Gaolin Fang, Jin Zhang, Hongbo Xu, Detect Events on Noisy Textual Datasets, In Proceedings of the 12th International Asia-Pacific Web Conference, Busan, Korea, April 2010.

[21] 宋丹、卫东、陈英. 基于改进向量空间模型的话题识别跟踪. 计算机技术与发展, 2006.

（下转第 51 页）